Impacts of Dropsonde Observations on Forecasts of Atmospheric Rivers and Associated Precipitation in the NCEP GFS and ECMWF IFS Models

LAUREL L. DEHAAN,[a] ANNA M. WILSON,[a] BRIAN KAWZENUK,[a] MINGHUA ZHENG,[a] LUCA DELLE MONACHE,[a] XINGREN WU,[b,c] DAVID A. LAVERS,[d] BRUCE INGLEBY,[d] VIJAY TALLAPRAGADA,[b] FLORIAN PAPPENBERGER,[d] AND F. MARTIN RALPH[a]

[a] *Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Deigo, San Diego, California*
[b] *NOAA/NWS/NCEP/EMC, College Park, Maryland*
[c] *Axiom Consultants, Inc., Rockville, Maryland*
[d] *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*

(Manuscript received 15 February 2023, in final form 26 September 2023, accepted 28 September 2023)

ABSTRACT: Atmospheric River Reconnaissance has held field campaigns during cool seasons since 2016. These campaigns have provided thousands of dropsonde data profiles, which are assimilated into multiple global operational numerical weather prediction models. Data denial experiments, conducted by running a parallel set of forecasts that exclude the dropsonde information, allow testing of the impact of the dropsonde data on model analyses and the subsequent forecasts. Here, we investigate the differences in skill between the control forecasts (with dropsonde data assimilated) and denial forecasts (without dropsonde data assimilated) in terms of both precipitation and integrated vapor transport (IVT) at multiple thresholds. The differences are considered in the times and locations where there is a reasonable expectation of influence of an intensive observation period (IOP). Results for 2019 and 2020 from both the European Centre for Medium-Range Weather Forecasts (ECMWF) model and the National Centers for Environmental Prediction (NCEP) global model show improvements with the added information from the dropsondes. In particular, significant improvements in the control forecast IVT generally occur in both models, especially at higher values. Significant improvements in the control forecast precipitation also generally occur in both models, but the improvements vary depending on the lead time and metrics used.

SIGNIFICANCE STATEMENT: Atmospheric River Reconnaissance is a program that uses targeted aircraft flights over the northeast Pacific to take measurements of meteorological fields. These data are then ingested into global weather models with the intent of improving the initial conditions and resulting forecasts along the U.S. West Coast. The impacts of these observations on two global numerical weather models were investigated to determine their influence on the forecasts. The integrated vapor transport, a measure of both wind and humidity, saw significant improvements in both models with the additional observations. Precipitation forecasts were also improved, but with differing results between the two models.

KEYWORDS: Atmospheric river; Dropsondes; Data assimilation; Numerical weather prediction/forecasting

1. Introduction

Operational numerical weather prediction (NWP) models can have inadequate representations of initial conditions in data-sparse areas, such as over the oceans, especially in cloudy conditions and precipitating areas associated with atmospheric rivers (ARs) and midlatitude cyclones (Zheng et al. 2021a). Since NWP is an initial value problem, improving the representation of the initial atmospheric state in NWP models has a strong potential to improve forecasts. This unsurprising finding has been consistently reported in the literature for decades (e.g., Arnold and Dey 1986; Rabier et al. 1996; Szunyogh et al. 2000; Cardinali 2009; Torn and Hakim 2009; Doyle et al. 2014; Ralph et al. 2014; Neiman et al. 2016; Demirdjian et al. 2020; among others). Specifically, Reynolds et al. (2019) showed that the initial condition errors in and near ARs are most effective in triggering short-term forecast errors for heavy precipitation events over the U.S. West Coast, based on landfalling ARs that brought record-breaking precipitation totals to California in early 2017. They also demonstrated that both kinematic and precipitation metrics are most sensitive to initial conditions in the mid- to lower-troposphere, corresponding to most data void regions for ARs (Zheng et al. 2021a).

Atmospheric River Reconnaissance (AR Recon), a program that uses targeted airborne and buoy observations over the Northeast Pacific Ocean, is an international research and operations partnership (RAOP) between academic institutions, government agencies, stakeholders, and operational centers to improve forecasting of ARs and their associated impacts in the western United States (Ralph et al. 2020; Cobb et al. 2023). AR Recon supports the improved prediction of landfalling ARs on the U.S. West Coast by supplementing global and regional observing systems with targeted dropsonde observations of

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-23-0025.s1.

Corresponding author: Laurel L. DeHaan, ldehaan@ucsd.edu

DOI: 10.1175/WAF-D-23-0025.1

atmospheric profiles of water vapor, temperature, and winds within ARs and in their vicinity as well as ocean surface drifters measuring sea level pressure among other parameters. Additional components of AR Recon include a partnership with the NOAA-funded Global Drifter Program led by the Scripps Lagrangian Drifter Laboratory to add barometers to the regular drifters (Centurioni et al. 2017). These are deployed annually with both the Air Force Reserve Command and ships of opportunity. AR Recon also supports airborne radio occultation, which provides complementary profiles of temperature and moisture and is currently under development for operational data assimilation (Haase et al. 2021). AR Recon flight tracks are designed to target sensitive regions where a better representation of the initial state of the atmosphere could most efficiently improve model accuracy in the verification regions. These sensitive regions are identified using fundamental physical knowledge of essential atmospheric structures like atmospheric rivers, vorticity strips, the upper-level jet, and extratropical cyclones, as well as using algorithms looking at ensemble sensitivities (Torn and Hakim 2008; Zheng et al. 2013) or applying adjoint model techniques (Doyle et al. 2014) to identify where observations to constrain the initial conditions of the models may be most useful. The major focus is on the core of the AR, located in the lower troposphere, below approximately the 700-hPa pressure level, and often associated with upper-level clouds, where it is difficult to observe with satellites (Zheng et al. 2021a). Secondary targets include dynamical features that may modulate the structure and evolution of the ARs. Daily forecast briefings include a comprehensive discussion of these targets in context of the forecast evolution and model differences (Ralph et al. 2020; Cordeira et al. 2017).

Recent studies have investigated the impacts of AR Recon dropsonde observations on the initial conditions and the forecast skill with a variety of analysis methods and NWP models. Lavers et al. (2018) identified large errors in the modeled integrated vapor transport (IVT) by comparing the differences between dropsonde observations from AR Recon 2018, the short-range forecasts (i.e., model background) and analyses in the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS). These IVT errors are largely attributable to wind errors at the top and above the planetary boundary layer, such as near the 850-hPa pressure level, where the data void is evident as well (Zheng et al. 2021a). Stone et al. (2020) assessed the impact of dropsondes collected during AR Recon 2018 by employing the forecast sensitivity observation impact (FSOI) computation available in the Navy Global Environmental Model (NAVGEM) system and found that the per observation impact from the dropsonde profiles can be more than double of that from the North American radiosonde observing network. Zheng et al. (2021b) systematically evaluated the impact of dropsondes taken during AR Recon 2016, 2018, and 2019 through full NWP data denial experiments using the regional Weather Research and Forecasting (WRF) Model (Skamarock et al. 2019) and the four-dimensional ensemble–variational (4DEnVar) method (Kleist and Ide 2015). They found significant beneficial impacts on the forecast skill of IVT and overall positive impacts on precipitation for lead times up to day 3. Follow-up work by Sun et al. (2022) further demonstrated the locally comparable impacts of assimilating dropsondes with that of assimilating satellite microwave radiances from the *NOAA-19* Advanced Microwave Sounding Unit A (AMSU-A), the *NOAA-19* Microwave Humidity Sounder (MHS), and the *Suomi National Polar-Orbiting Partnership* (*SNPP*) Advanced Technology Microwave Sounder (ATMS) through both the FSOI and full NWP experiments with the WRF and 3DVar framework. Positive impacts from data collected during AR Recon 2020 on the forecast skill in the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) system are shown in Lord et al. (2023a,b).

This manuscript investigates the impact of AR Recon data on the forecast skill over the U.S. West Coast in two major global operational NWP systems. Cycled data denial experiments were performed in both operational modeling systems by withholding dropsonde observations during the data assimilation step from 22 intensive observation periods (IOPs) during AR Recon 2019 and 2020. This work is different from previous studies in that a systematic evaluation will be conducted in multimonth simulations for two seasons with two leading cycled global NWP systems (i.e., the GFS at NCEP and the IFS at ECMWF). In particular, this study will focus on the impact on IVT as the key parameter representing AR features and precipitation as a critical metric for water resource management. Section 2 of the manuscript provides details on the data and methods used in this work. Section 3 describes the results of the comparison in forecast accuracy between the control forecasts including dropsonde data and the denial forecasts excluding those data. Section 4 provides a discussion and summary of these results in the context of ongoing expansions and improvements in the AR Recon program, and a brief conclusion is in section 5.

2. Data and methods

a. *Models*

This study utilizes two NWP models: the IFS (Hersbach et al. 2020) from ECMWF and the GFS (Yang and Tallapragada 2018) from NCEP. Both of these models provided control and denial runs for all 22 IOPs in 2019 and 2020. For this study we utilize the forecasts from 24 to 120 h from each of the models, initialized at 0000 UTC. There are numerous differences between the models and their data assimilation systems. Consequently, this study is not intended to be a comparison between the two models, but rather a presentation of two independent assessments of the impacts of dropsondes on deterministic forecasts in leading global operational NWP models. The following subsections describe the model details and the experiments designed for this study.

1) ECMWF IFS

The ECMWF IFS version cycle 45r1 (https://www.ecmwf.int/en/about/media-centre/news/2018/ifs-upgrade-improves-extended-range-weather-forecasts) was used for control and denial experiments for the period from 0000 UTC 1 February 2019 to 0000 UTC 28 February 2019 and cycle 46r1 ran experiments

TABLE 1. Number of dropsondes assimilated into each of the models for each IOP during the 0000 UTC window. Dates are 0000 UTC and represent the time the IOP was centered on.

IOP date	No. of drops NCEP/ECMWF	IOP date	No. of drops NCEP/ECMWF	IOP date	No. of drops NCEP/ECMWF
2 Feb 2019	52/52	24 Jan 2020	37/32	16 Feb 2020	30/27
11 Feb 2019	24/21	29 Jan 2020	21/23	21 Feb 2020	29/26
13 Feb 2019	47/43	31 Jan 2020	24/22	24 Feb 2020	55/51
24 Feb 2019	45/43	4 Feb 2020	70/65	2 Mar 2020	74/72
26 Feb 2019	35/33	5 Feb 2020	30/30	7 Mar 2020	30/26
1 Mar 2019	59/50	6 Feb 2020	59/54	8 Mar 2020	54/49
		14 Feb 2020	27/28	9 Mar 2020	29/26
		15 Feb 2020	78/73	10 Mar 2020	39/51

from 0000 UTC 24 January 2020 to 0000 UTC 18 March 2020 for control and denial experiments, for a total of 83 days. The horizontal grid spacing for this application was approximately 0.22° or 25 km. The model has 137 vertical layers with a model top of 0.01 hPa. Data assimilation was performed using 4D-Var (Rabier et al. 2000) with 12-h windows (2101–0900 UTC for 0000 UTC analysis time). IFS cycle 45r1 included treatment of sonde drift where positions of each level are in the binary report (Ingleby et al. 2018), and cycle 46r1 extended that treatment to binary dropsonde reports (Ingleby et al. 2019), although for dropsondes the impact of drift is generally small. More details of in situ observations and their use are found in Pauley and Ingleby (2022).

In operations, the ECMWF model assimilated data from AR Recon dropsondes in its deterministic forecasts, which are referred to as the "control forecasts" in this study. The denial run—referred to as the "denial forecasts"—was carried out the same way as in the control run except rejecting AR Recon dropsonde data from the operational observation dataset. Model outputs interpolated to 0.25° horizontal resolution were employed to analyze the difference (i.e., impacts from dropsonde data) between these two sets of forecasts.

2) NCEP GFS

Data denial experiments were also conducted with the deterministic NCEP GFS version 15 (GFSv15) global model for the period from 0000 UTC 1 February 2019 to 0000 UTC 8 March 2019 and from 0000 UTC 24 January 2020 to 0000 UTC 18 March 2020, for a total of 91 days. The Geophysical Fluid Dynamics Laboratory (GFDL) finite-volume cubed-sphere dynamical core (Lin 2004; Putman and Lin 2007; Harris and Lin 2013; Harris et al. 2020) and a suite of physical parameterizations comprise the GFS model. The GFSv15 upgraded physical parameterization package includes replacement of Zhao–Carr microphysics with the more advanced GFDL microphysics (Zhou et al. 2019). In the operational setting, the model was run with a horizontal resolution of approximately 13 km. There are 64 levels in the vertical with the model top at 0.2 hPa. The denial forecasts were carried out the same way as in the control except rejecting AR Recon dropsonde data from the observation dataset assimilated in the control forecasts. Note that while the 2020 upgraded GFSv15 was made operational on 4 March 2020, the code was implemented for these experiments before it was operational. As such, this GFS is a consistent version of the assimilation and model throughout both the control and denial experiments for each year.

The corresponding Global Data Assimilation System (GDAS) at NCEP generated model analyses with the hybrid 4DEnVar method (Kleist and Ide 2015; Wang and Lei 2014). The hybrid algorithm combines uncertainty estimates from the ensemble and a variance that is derived from model data, constant in time, but spatially varying over the globe. The flow-dependent part of the background error-covariance matrix was based on the 80-member ensemble created by the ensemble Kalman filter method. In GFSv15, the ensemble data were at a horizontal grid spacing of approximately 25 km. The system was cycled every 6 h centered at 0000, 0600, 1200, and 1800 UTC.

While both models also assimilate satellite data, it is relevant to note that the ECMWF model assimilates more satellite radiance data than the NCEP model (Geer et al. 2021). Such differences are expected to affect differences in the model background, the innovation (i.e., the absolute difference between the observed and simulated variable) from dropsondes, and the dropsonde data impacts.

b. *AR Recon 2019/20*

During the 2019 and 2020 seasons there were 22 successful IOPs (6 in 2019 and 16 in 2020) each utilizing between one and three aircraft, that deployed dropsondes assimilated into the NCEP and ECMWF forecasts (Table 1). In addition to dropsondes assimilated in the 0000 UTC window, listed in Table 1, the NCEP model assimilated an additional 10 dropsondes at other cycle times, and the ECMWF model assimilated an additional 83 dropsondes over all the IOPs in 2019 and 2020. The difference in the number of dropsondes assimilated between the two models is partly due to the duplicate checks in the individual systems and the different lengths of data assimilation windows. There are also cases where some reports reached one center but not the other. Maps of synoptic-scale conditions, including IVT values, along with the location of the drops are available for each IOP at https://cw3e.ucsd.edu/arrecon_data/.

The aircraft used to deploy dropsondes for these field campaigns include the U.S. Air Force (USAF) Reserve Command WC-130Js and a NOAA G-IV. Dropsondes manufactured by Vaisala (RD41) were released from these aircraft in and around ARs from an altitude of approximately 30 000 feet (approximately 9100 m) for the WC-130Js and 40 000 feet

(approximately 12 200 m) for the G-IV. Each dropsonde records pressure, temperature, and relative humidity as it descends through the atmosphere, with wind and height also reported via GPS (Cobb et al. 2023). The USAF reports were in alphanumeric format, while some of the NOAA profiles were also in binary format, providing more levels and drift information. These data provide full profiles of critical parameters in regions specifically targeted during the AR Recon flight planning procedures.

c. Variables and metrics

A major goal of AR recon is to improve forecasts of ARs and precipitation. Consequently, this study compares the control forecasts (that assimilate the dropsondes) and the denial forecasts (that exclude the dropsondes) both in terms of integrated vapor transport (IVT) and 24-h accumulated precipitation.

IVT is calculated as

$$\mathrm{IVT} = \frac{1}{g}\int_{P\mathrm{sfc}}^{P_{\mathrm{top}}} q\mathbf{V}\, dp,$$

where g is the gravitational acceleration (m s^{-2}), q is the specific humidity (kg kg^{-1}), and $\mathbf{V}$ is the horizontal vector wind (m s^{-1}). For ECMWF Reanalysis v5 (ERA5; Hersbach et al. 2020) and ECMWF IFS, IVT is calculated from 1000 hPa (P_{sfc}) to 300 hPa (P_{top}). For NCEP it is calculated from 1000 to 200 hPa. Although the top pressure differs, Ralph et al. (2017) suggests IVT is insensitive to additional data at pressures lower than 300 hPa.

IVT is verified using ERA5 (Hersbach et al. 2020) as a proxy to observations with computations made on ERA5's 0.25° grid, and precipitation is verified using the NCEP Stage-IV quantitative precipitation estimate (Lin and Mitchell 2005) as truth, with computations made on Stage-IV's 4-km grid. Regridding to put IVT on the ERA5 grid is done with the nearest neighbor method, and regridding to put precipitation on the Stage-IV grid is done with the budget method. All regridding was done using the Model Evaluation Tools (MET) package (https://www.dtcenter.org/community-code/model-evaluation-tools-met). While no verification dataset is perfect, Stage-IV data are widely used as a reference for precipitation because of its accuracy (e.g., Beck et al. 2019). Note that Stage-IV data are only available over land for CONUS, which limits the regions of precipitation verification. Likewise, ERA5 has been shown to have the smallest errors in IVT at the location of dropsondes compared to two other reanalysis datasets (Cobb et al. 2021). The ERA5 reanalysis DA system assimilates the wind, temperature, and humidity data from AR Recon dropsondes when they are available, as do both the ECMWF and NCEP operational systems. These facts are important to consider when interpreting the dropsonde impact results from model runs.

For both precipitation and IVT, mean absolute error (MAE) and standard point-to-point spatial correlation are used as verification metrics with thresholds of 13, 25, and 50 mm $(24\ \mathrm{h})^{-1}$ for precipitation, and 250 and 500 kg m^{-1} s^{-1} for IVT. Note that the precipitation thresholds are for a 24-h accumulation ending at 0000 UTC on the valid day. These are common thresholds for precipitation and IVT verification (e.g., Cordeira et al. 2017; DeHaan et al. 2021, among many others), and the IVT thresholds also correspond to the AR scale (Ralph et al. 2019). Two additional metrics are used for precipitation: fractions skill score (FSS; Roberts and Lean 2008) and watershed intensity error. FSS is a neighborhood verification method that compares the fraction of grid boxes above a threshold between the forecast and observation for a given number of grid boxes. The FSS ranges in value from 0 (no skill) to 1 (perfect forecast). Since the models were regridded to a finer resolution before performing this analysis, a neighborhood size that is smaller than either the ECMWF or NCEP initial grid resolution would provide little additional information. We have chosen to use a 9 × 9 square (36 km × 36 km) as well as a 15 × 15 square (60 km × 60 km) for the FSS verification. The former is the minimum reasonable size given the resolution of the models and the latter is representative of other larger tested neighborhood sizes, from 52 km × 52 km to 76 km × 76 km.

Given the importance of extreme precipitation skill at the watershed level to water resource management, we also consider watershed intensity error, as defined below. For this study, we focus on three California watersheds: the Russian, Yuba/Feather, and Santa Ana River watersheds (Fig. 1d), all of which are currently being assessed for Forecast Informed Reservoir Operations viability (e.g., Jasperse et al. 2020). Since the concern is for the upper extremes of precipitation, we created a variation of mean areal precipitation (MAP), which we will refer to as watershed intensity. Watershed intensity is computed by finding the 90th percentile value of precipitation of all grid points in each watershed for a given time (rather than the mean, as done for MAP), which focuses on the grid points with the most precipitation each day. This initial calculation results in an intensity value for each forecast and each watershed, and an error, which is the difference between the forecast and Stage-IV 90th percentile values. To then focus on the days with more extreme precipitation, we only use days where the Stage-IV 90th percentile value is above a threshold. For the Russian and Yuba/Feather watersheds the threshold is 50 mm $(24\ \mathrm{h})^{-1}$. For the Santa Ana watershed the threshold is 20 mm $(24\ \mathrm{h})^{-1}$. These thresholds roughly correspond to the wettest 25% of days in the time period considered in this study. For comparison, when the 90th percentile value is 50 mm $(24\ \mathrm{h})^{-1}$ in the Russian or Yuba/Feather watersheds, the MAP averages approximately 20 mm $(24\ \mathrm{h})^{-1}$, and when the 90th percentile value is 20 mm $(24\ \mathrm{h})^{-1}$ in the Santa Ana watershed, the MAP is approximately 8 mm $(24\ \mathrm{h})^{-1}$.

d. Spatial domain, valid days, and lead times

The domains considered for verification are shown in Fig. 1 (approximately 115°–170°W and 18°–56°N for IVT and land areas west of 115°W for precipitation). For each threshold and time, the only grid points within the full domain that are used in the verification are points where at least one of the observation, control forecast, or denial forecast has values above the threshold. These limited domains for the example shown in Fig. 1 are indicated by the heavy black line in Fig. 1a (IVT) and Fig. 1c (precipitation). To focus on areas where the dropsondes are most likely to have an effect, the domain used for

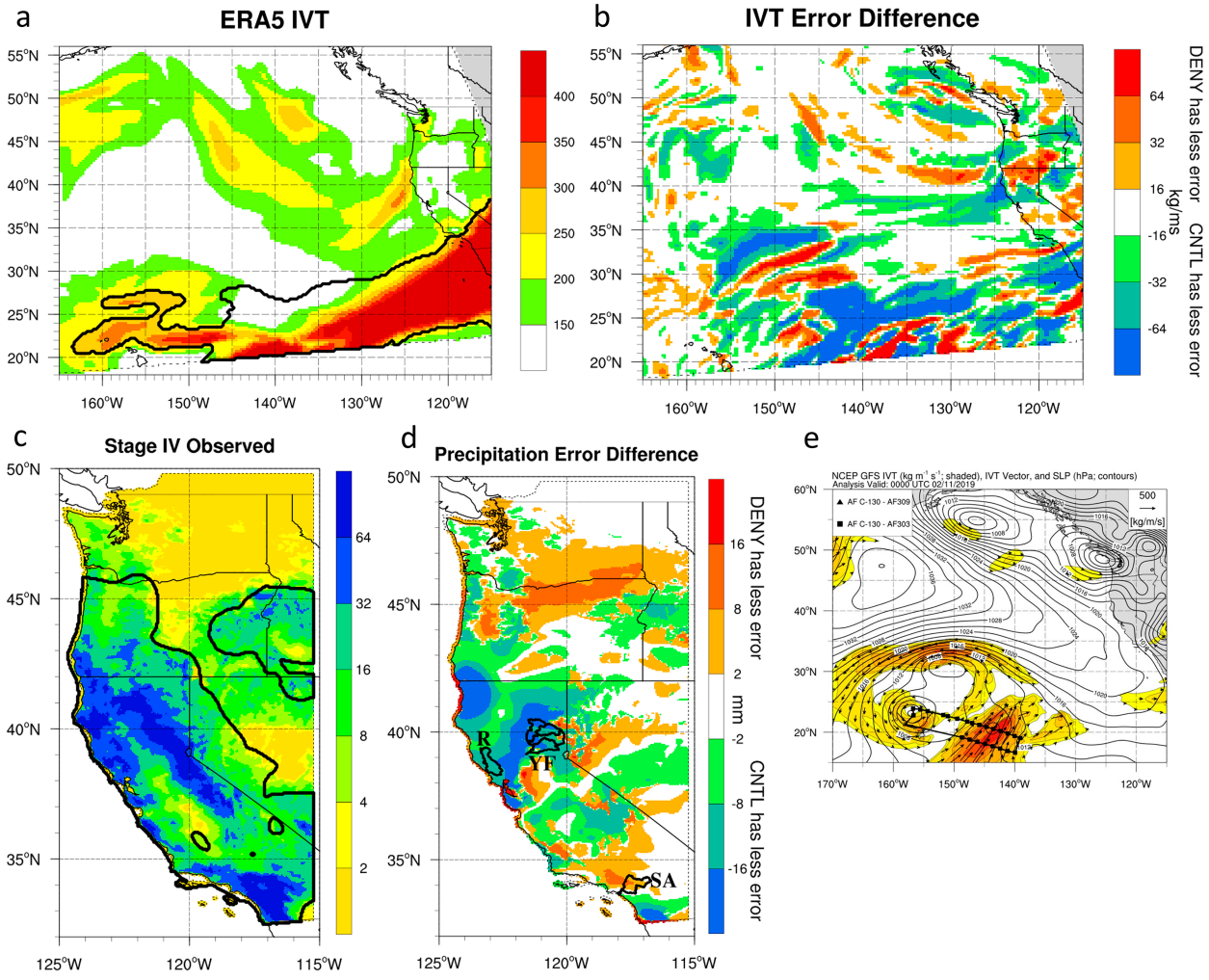

FIG. 1. Observed values and error differences for IVT and precipitation, valid on 15 Feb 2019. (a) ERA5 IVT (color shading; kg m^{-1} s^{-1}) at 0000 UTC 15 Feb 2019; the heavy black line indicates the area used for IVT verification in this example (union of control, denial, and ERA5 above 250 kg m^{-1} s^{-1}); (b) difference in IVT error (MAE; color shading; kg m^{-1} s^{-1}) between control and denial (control − denial) for the 96-h ECMWF forecast valid at 0000 UTC 15 Feb 2019 (blue indicates that the control has less error); (c) Stage-IV 24-h accumulated precipitation ending at 0000 UTC 15 Feb 2019; the heavy black line indicates the area used for precipitation verification in this example (the union of control, denial, and Stage-IV above 13 mm); (d) difference in 24-h accumulated precipitation error (color shading; mm) between control and denial (control − denial) for the 96-h ECMWF forecast ending at 0000 UTC 15 Feb 2019. Three watersheds are outlined: Russian (R), Yuba/Feather (YF), and Santa Ana (SA). (e) NCEP GFS analysis IVT (color shading and vectors) and mean sea level pressure valid at 0000 UTC 11 Feb 2019, and locations of dropsondes during 2019 IOP2.

IVT verification is limited to ARs east of the western most dropsonde from the most recent IOP.

Valid days and lead times are limited to times when there is a reasonable expectation of the dropsondes influencing the forecast. For IVT, starting on an IOP date, each consecutive day with an AR east of the western most dropsonde is used as a valid day in the analysis. For the sake of determining valid days, ARs are defined from IVT using Method for Object-Based Diagnostic Evaluation (MODE; Bullock et al. 2016) which is a part of the MET package. ARs are required to have a minimum threshold of 250 kg m^{-1} s^{-1} and a minimum length of 2000 km (DeHaan et al. 2021). Similarly, valid days for precipitation require a region of at least 400 km^2 with 24-h accumulated precipitation of 13 mm or greater.

Forecast initialization times are limited to forecasts initialized from 1 to 5 days after an IOP. While the choice of 5 days was subjective, a cursory look at all the events in this analysis indicated that both the AR that was sampled in an IOP and the precipitation appearing to result from that AR were completed within 5 days. In addition, AR Recon is focused on improving forecast skill for lead times of 1–5 days.

The reason for choosing the spatial and temporal limitations defined in this section is to focus on the regions and times specifically targeted for improvement in AR Recon. Future work may investigate the success of AR Recon in regions not specifically

TABLE 2. Comparison of validation metrics in the control and denial forecasts with the ECMWF model for the 96-h forecast, valid at 0000 UTC 15 Feb 2019. Precipitation is the 24-h accumulation ending at 0000 UTC 15 Feb 2019. The verification area is indicated by the heavy black lines in Figs. 1a and 1c.

	IVT control	IVT denial	Precip control	Precip denial
MAE [25 mm (24 h)$^{-1}$ or 250 kg m^{-1} s^{-1}]	142 kg m^{-1} s^{-1}	173 kg m^{-1} s^{-1}	16 mm (24 h)$^{-1}$	29 mm (24 h)$^{-1}$
Spatial correlation [25 mm (24 h)$^{-1}$ or 250 kg m^{-1} s^{-1}]	0.63	0.43	0.48	0.36
FSS [25 mm (24 h)$^{-1}$]			0.77	0.59
Watershed intensity error (average of 3 watersheds)			73 mm (24 h)$^{-1}$	80 mm (24 h)$^{-1}$

targeted by the flights or assess sensitivity to perturbations in the domains chosen here. While no formal sensitivity test was performed, anecdotal evidence suggests the results are not sensitive to modest changes in the full domain, and the days excluded generally did not have significant differences between the control and denial forecasts.

Initially, the selection of valid days and lead times is the same for each threshold. At the higher thresholds, we additionally exclude valid days that do not have enough grid points above the threshold to reliably calculate the spatial prediction comparison test (SPCT; Hering and Genton 2011; Gilleland 2013). For both precipitation and IVT, the minimum size was approximately 100 grid points.

e. Significance

A major aim of this study is to look at the statistical significance of the differences between the control and denial forecasts. SPCT is used to test for significant differences in MAE and spatial correlation. A key advantage of using the SPCT test statistic is that it allows a test of significance for the differences in a pair of forecasts at a single time. The SPCT uses MAE and correlation as loss functions and accounts for spatial correlation of the grid points by using an exponential variogram in the computation of the test statistic. The variogram is a function which describes the degree of spatial dependence in a field.

To determine significance when considering multiple forecasts, confidence intervals are computed using Matlab's bootstrapping function (bootci) with a sample size of 1000. For both, we define significance at 90%.

3. Results

a. Examples from February 2019

One representative case study from this work was an AR that made landfall over California on 13–15 February 2019 (Fig. 1). This storm was categorized as an AR 5 (extreme or exceptional) using the AR scale of Ralph et al. (2019) and caused widespread impacts across California including landslides, flash flooding and evacuations (Hatchett et al. 2020). This AR was sampled during two IOPs on 11 and 13 February 2019. A total of 26 dropsondes were released by one aircraft to the east of Hawaii sampling the tropical moisture export (Chen et al. 2022) and the developing AR on 11 February 2019 (Fig. 1e). Two days later, 53 dropsondes were released by two aircraft over the northeast Pacific sampling the AR shortly before it made landfall over Northern California. For the purpose of illustrating the metrics and analysis used in this study, we focus on the ECMWF 96-h forecast initialized on 11 February. We consider IVT MAE and correlation using a 250 kg m^{-1} s^{-1} threshold and precipitation MAE and correlation using a 25 mm (24 h)$^{-1}$ threshold, as well as FSS and watershed intensity error (difference in MAE: Figs. 1b,d). The results for these metrics at these thresholds suggest that the dropsondes had a positive influence on both the forecasted IVT and precipitation (Table 2). For example, with the control forecast, the IVT MAE is reduced by 31 kg m^{-1} s^{-1} compared to the denial forecast. Similarly, the precipitation MAE is reduced by 13 mm (24 h)$^{-1}$ when the dropsondes are included. For this case, every metric was improved with the use of the dropsondes, and SPCT showed that the improvements in both the MAE and correlation, for both precipitation and IVT are all significant at 90%.

However, the influence of dropsondes on the forecast is not always so clear. The ECMWF forecast initialized at 0000 UTC 12 February 2019, shows mixed results (Fig. 2). This forecast was initialized a day after the IOP on 11 February. The control has less IVT MAE than the denial at 24 h (65 versus 72 kg m^{-1} s^{-1}; valid 0000 UTC 13 February, Fig. 2g), and the control also has less precipitation MAE for the accumulation from 0000 UTC 13 February to 0000 UTC 14 February (12 versus 14 mm; the 48-h forecast). Conversely, the denial has less IVT MAE at 48 h (87 versus 93 kg m^{-1} s^{-1}; Fig. 2h), and less precipitation MAE for the accumulation between 48 and 72 h (14 mm versus 15 mm; valid at 0000 UTC 15 February 2019). In this case it is quite possible that the denial out-performs the control in the later lead times since the influences of the dropsondes are decreasing with time. While this was not the only case with a reduction of positive dropsonde influence with time, it was not found to happen in general. This points to the complicated influence of dropsondes on individual forecasts and the need to look more broadly over many forecasts.

b. Results across all IOPs

The metrics computed in the case study were repeated for all the valid days and lead times in which an impact from dropsondes was most likely to be seen, as described above. The shaded boxes in Figs. 3a and 4a show the valid days and lead times used for IVT, and the same is shown for precipitation in Figs. 3b and 4b, for the ECMWF and NCEP models, respectively. We will refer to each valid day and lead time (i.e., each shaded box in Fig. 3 or Fig. 4) as an "instance" throughout the manuscript. The values in the grid boxes in Figs. 3 and 4 are the differences in MAE between the control

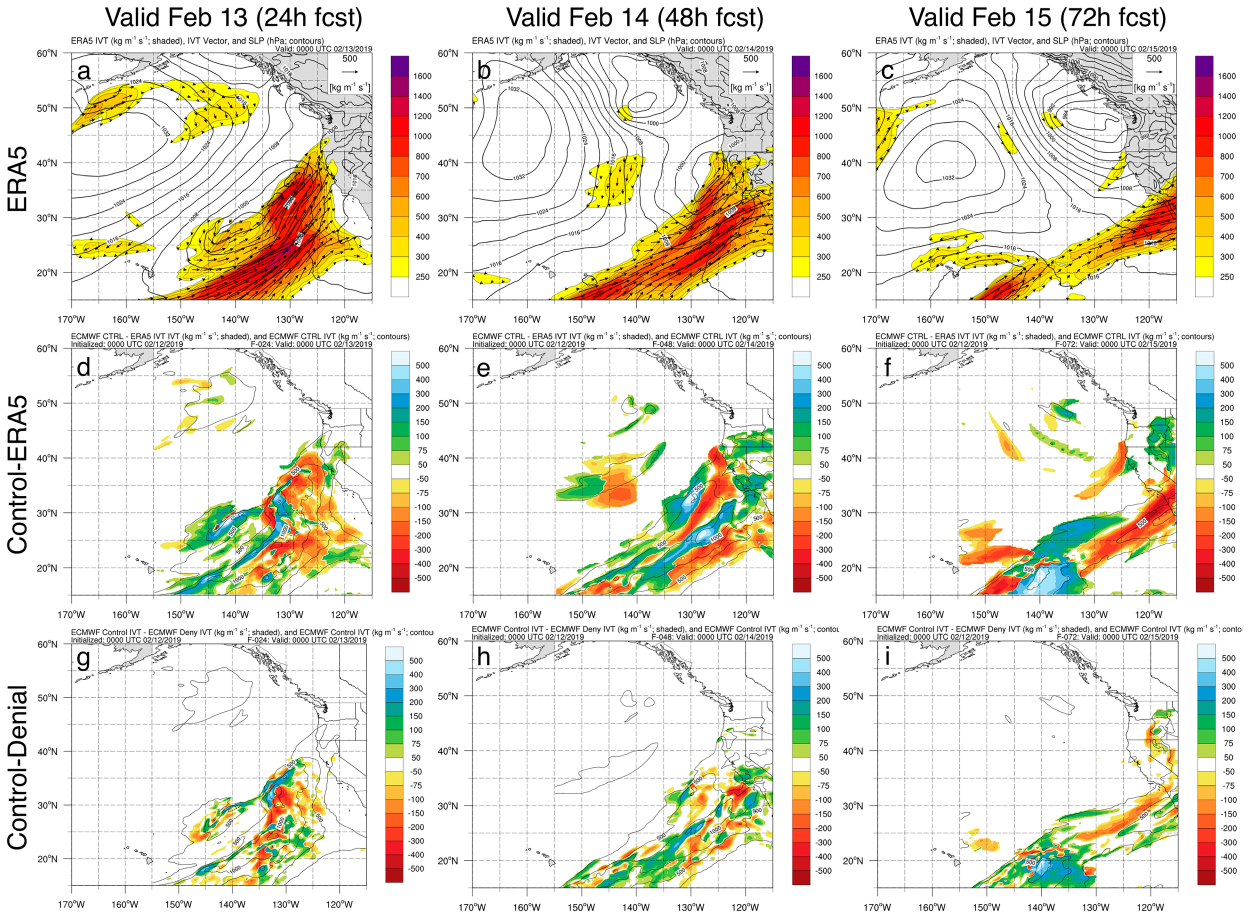

FIG. 2. IVT fields and IVT forecast differences for the 13–15 Feb 2019 case. (a)–(c) IVT (color shading; kg m^{-1} s^{-1}) and sea level pressure (contours; hPa) from ERA5, valid at 0000 UTC 13–15 Feb 2019. (d)–(f) Difference between ECMWF Control forecast and ERA5 IVT (control − ERA5), for the forecast initialized at 0000 UTC 12 Feb 2019, valid on 13–15 Feb. (g)–(i) Difference between the ECMWF control and denial (control − denial) forecasts for the forecast initialized at 0000 UTC 12 Feb 2019, valid on 13–15 Feb.

and denial forecasts, with significant differences based on SPCT shaded red or blue. Cells with no values are valid days and lead times that are excluded from the analysis based on the criteria listed above. Similar tables were created for other thresholds for MAE as well as all thresholds for the correlation metric and are shown in the supplemental material.

For the ECMWF model (Fig. 3) there are approximately twice as many instances with significant improvements (blue boxes) than there are instances with significant degradations (red boxes) from the assimilation of the dropsondes for both precipitation at 25 mm $(24\ h)^{-1}$ and IVT at 250 kg m^{-1} s^{-1}. Most improvements in the precipitation forecasts are at lead times between 48 and 120 h, while dropsonde impacts on the 24-h precipitation forecasts are overall neutral or slightly negative. It is quite possible that more improvement would be seen at the 24-h lead time if the precipitation verification included values over the ocean instead of being limited to land values, which are far from the dropsonde locations. However, the focus of AR Recon precipitation forecast improvement is over land and is therefore the focus of this work.

The NCEP model also showed improvements from the assimilation of the dropsonde data at several valid days and lead times (Fig. 4); however, instances where there are significant improvements with the NCEP control forecast differ from the ECMWF model. For IVT (Fig. 4a), the NCEP model, like ECMWF, has approximately twice as many instances when the control forecast significantly outperforms the denial forecast, as compared to instances that have degradations with the control forecast. For 25 mm $(24\ h)^{-1}$ precipitation there are fewer instances with the NCEP model that show significant improvement (based on SPCT) with the dropsondes assimilated. However, as will be seen later, there are many cases that show improvement, but do not pass the SPCT significance test.

One method to summarize the differences between the control and denial forecasts is to count the number of instances (valid days and lead times) where one is significantly better than the other based on the SPCT and compare the mean improvement from those instances (Fig. 5). For example, there are 19 instances where the control forecast outperforms the denial forecast in IVT MAE for ECMWF, considering both the 250 and 500 kg m^{-1} s^{-1} thresholds, while there are only

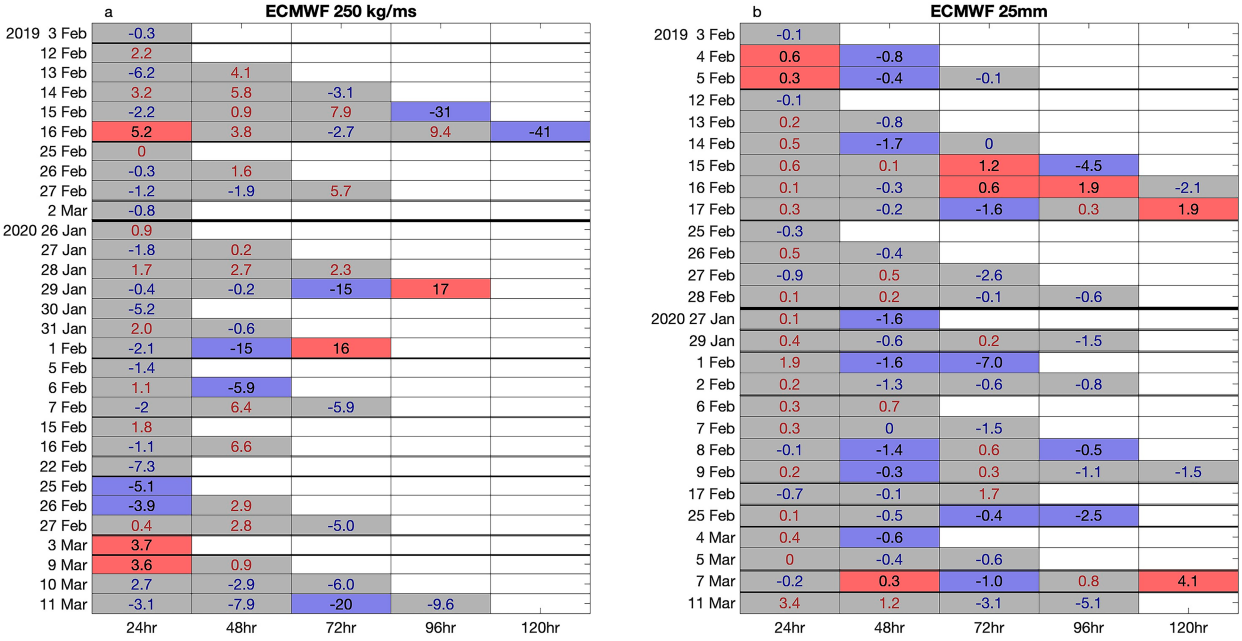

FIG. 3. Table of differences (control − denial) in MAE from the ECMWF model for valid days and lead times included in the analysis (all shaded cells) for (a) IVT at 250 kg m^{-1} s^{-1} and (b) precipitation at 25 mm (24 h)$^{-1}$. Blues (numbers or shading) indicate control has less error; reds indicate denial has less error. Cells are shaded red or blue when the difference between control and denial MAE is significant at 90% using SCPT.

8 instances where the denial forecast outperforms the control (Fig. 5a). Eight of the instances that were improved with dropsondes occurred at a threshold of 250 kg m^{-1} s^{-1} (blue cells in Fig. 3a) and the other 11 occurred at the 500 kg m^{-1} s^{-1} threshold (Fig. 1 in the online supplemental material). Note that there are 5 forecasts that were improved with the control at both 250 and 500 kg m^{-1} s^{-1}, and these are counted separately. In addition, the average reduction in MAE with the control run in the 19 instances when it had less error is greater than the average reduction in the 8 instances when the denial

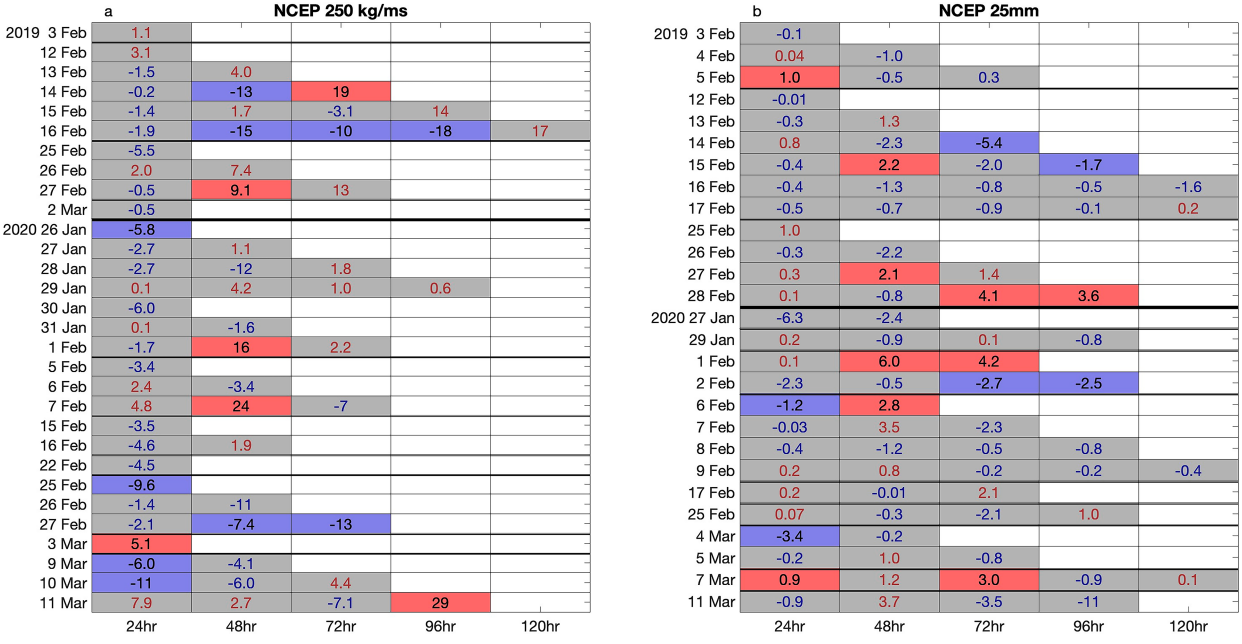

FIG. 4. As in Fig. 3, but for NCEP.

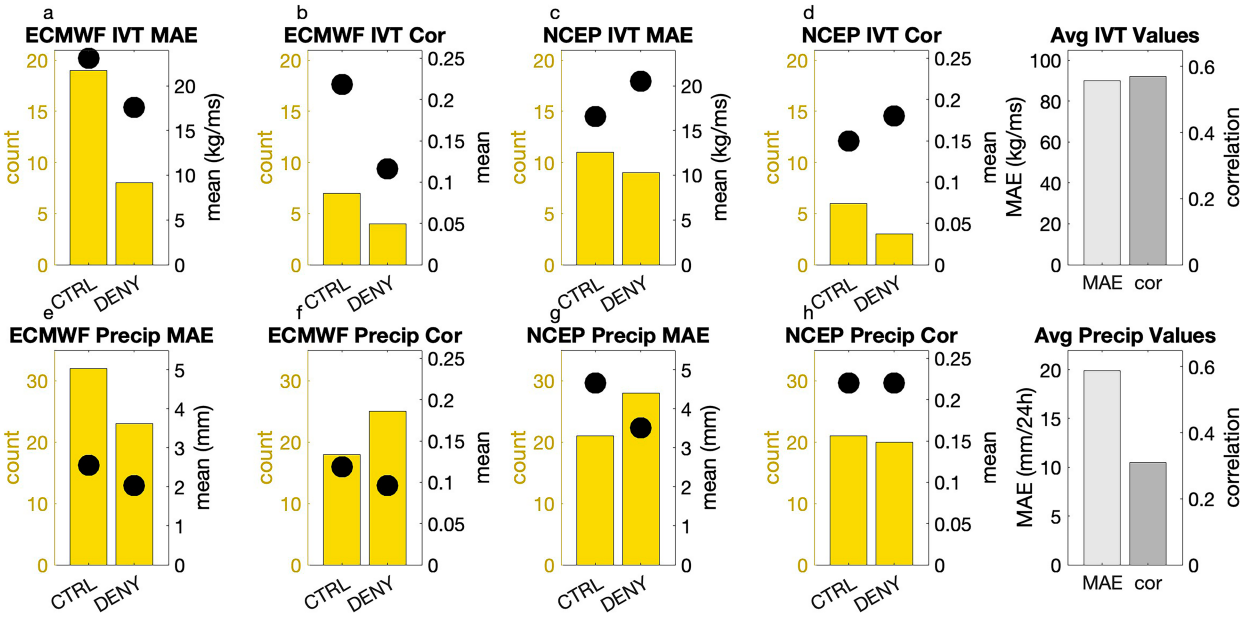

FIG. 5. Summary of the significant results from Figs. 2 and 3 together with supplemental Figs. 1–8. (a)–(h) Yellow bars are the count of instances (valid days and lead times shown as blue or red shaded cells in Figs. 2 and 3 and supplemental Figs. 1–8) summed over all thresholds [250 and 500 kg m^{-1} s^{-1} for IVT, and 13, 25, and 50 mm $(24\ h)^{-1}$ for precipitation] where either the control or denial significantly outperforms the other based on SPCT (as indicated by the label). Black dots are the mean improvement of those instances. Right-hand plots with gray bars show the average values (over all significant instances, from both models, and control and denial) of MAE and correlation for reference.

run has less error. The situation is similar for IVT with the correlation metric (Fig. 5b), where the control run has more instances with a higher correlation and a larger improvement in correlation compared to the denial run. For precipitation MAE (Fig. 5e) the sense is the same, while precipitation correlation (Fig. 5f) shows more cases when the denial outperforms the control, albeit with a smaller mean improvement than the control. For both precipitation metrics, excluding the 24-h forecasts paints a different picture. With MAE, 7 of the 23 instances of the denial outperforming the control happen at 24-h, while only 1 of the 32 instances of control outperforming denial occur at a 24-h lead time. For correlation, over half (13 of 25) of the instances of the denial forecast outperforming the control occur at 24 h, while only 1 of 18 occur at 24-h for the control. This highlights the positive impact that the dropsondes had on forecast accuracy for precipitation at lead times greater than 24 h for the ECMWF model.

Results for IVT MAE and correlation for the NCEP model (Figs. 5c,d) show that there are more instances when the control forecast outperforms the denial forecast than vice versa, while the means of those instances are smaller with the control. For precipitation correlation, the control and denial forecasts are quite similar. Precipitation MAE shows more instances where the denial outperforms the control forecasts; however, the mean improvement when the control has a better forecast is larger than when the denial is the better forecast, which overall represents similar MAE between the two. It should be noted that using SPCT significance as a precursor to skill analysis focuses on improvements for single events but excludes consistent smaller differences over the seasons that are not individually significant.

Complementing the previous analysis, which was limited to initializations and lead times significant with SPCT, we also performed analysis utilizing all initializations and lead times indicated in Figs. 3 and 4 and determined significance based on bootstrapping. We first consider averaging over thresholds and separating the lead times (Fig. 6). For IVT, both models show generally modest improvement in MAE at all lead times, with the exception of ECMWF at 48 h. Both models show stronger results using the IVT correlation metric, with the ECMWF model showing significant improvement at 24 and 72 h, and the NCEP model showing significant improvement at all three lead times (Figs. 6b,f). Note that longer lead times are not shown here given the small number of valid days considered at those lead times (Table 3). For precipitation, the ECMWF model has significant improvements in MAE with the control forecasts at lead times of 48, 72, and 96 h (Fig. 6c). The NCEP model has significant improvements at lead times of 24 and 72 h (Fig. 6g). The correlation of precipitation has mixed results, with the only significant improvements with the control forecasts occurring at 48- and 72-h lead times for the NCEP model (Figs. 6d,h).

Next, we consider averaging over lead times and separating the thresholds (Fig. 7). Both models show improvements in IVT MAE (Figs. 7a,e) and correlation (Figs. 7b,f) at both the 250 and 500 kg m^{-1} s^{-1} thresholds, with significant improvements in IVT correlation using the 500 kg m^{-1} s^{-1} threshold. There are mixed results for the precipitation correlations (Figs. 7d,h). However,

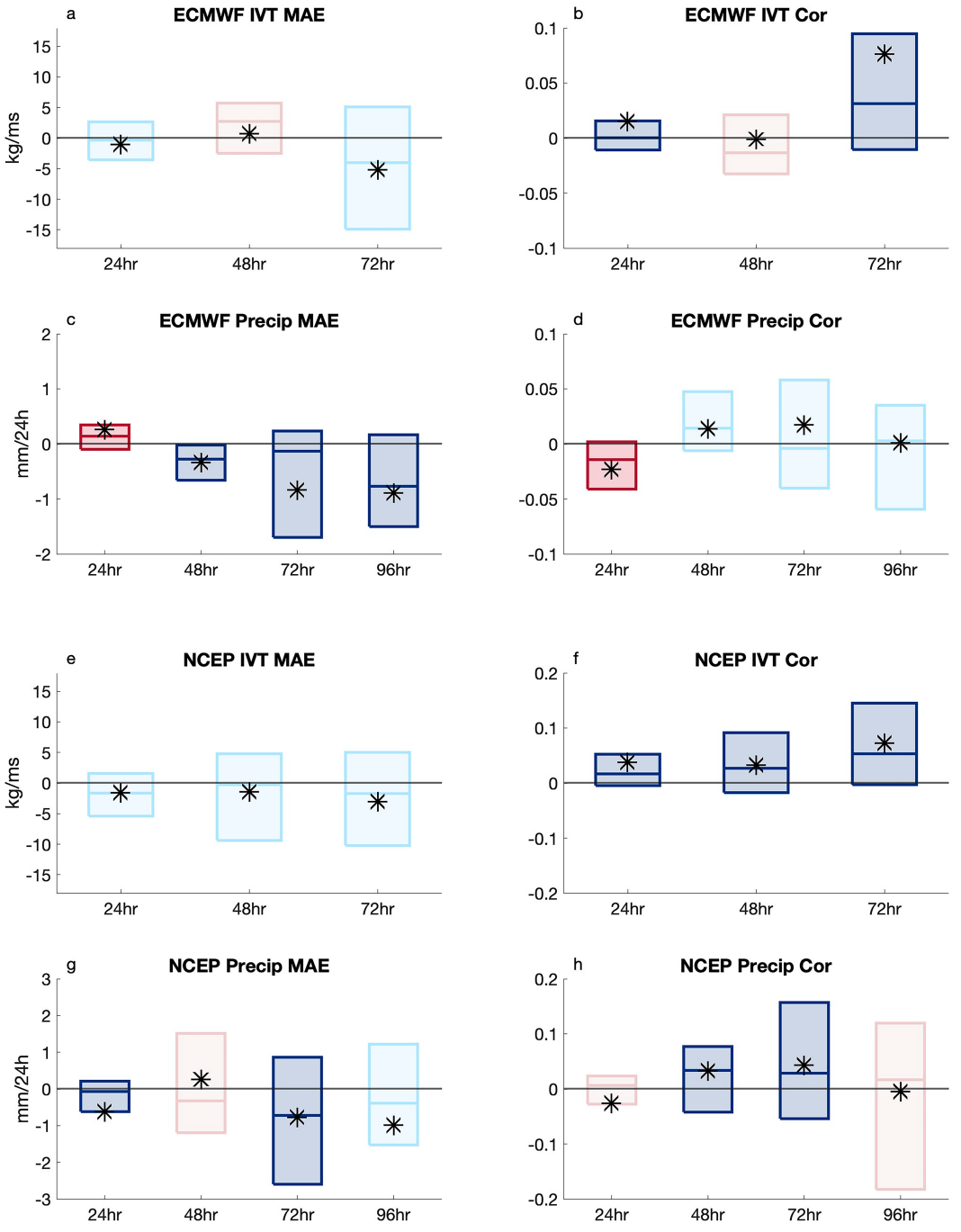

FIG. 6. Averages of differences (control − denial) in error or correlation across all thresholds. Boxes are the interquartile range; the middle line is the median and the asterisk shows the mean. Blue colors indicate the control has less MAE or higher correlation in the mean; red colors indicate the denial has less MAE or higher correlation in the mean. Darker shades indicate significant differences in the mean based on a 90% confidence interval computed with bootstrapping.

there is consistent improvement in the control precipitation forecast as determined by MAE for both models, with significant improvement with the ECMWF model at all thresholds (Fig. 7c), and significant improvements with the NCEP model at 50 mm $(24\ \text{h})^{-1}$ (Fig. 7g).

To better understand the differences in precipitation skill, we look at two metrics that are not point-based. Point-based metrics are sensitive to model resolution and can add the so-called "double penalty" to models with a finer resolution (Gilleland et al. 2009). To address these issues fractions skill

TABLE 3. The number of valid days at each lead time ("instances") in which the forecasts meet the predefined criteria.

	24 h	48 h	72 h	96 h	120 h
Precip	27	24	17	11	4
IVT	29	19	11	4	1

scores (Roberts and Lean 2008) were calculated using a 9 × 9 grid (36 km × 36 km) and a 15 × 15 grid (60 km × 60 km) (Fig. 8). For reference, the average value (over models, lead times, valid days, control, and denial) of FSS was approximately 0.47 for the method with the smaller neighborhood, and 0.51 for the larger neighborhood method. As seen with other metrics for the ECMWF model, there is a reduction in skill at the 24-h lead time with the control forecast that affects all thresholds. Using the smaller neighborhood size there is an improvement at all other lead times, while the larger neighborhood size does not show improvement until after the 48-h lead time. NCEP has improvements with the control forecasts at all lead times, with significant improvements at the shorter lead times and higher thresholds, and general agreement between the two neighborhood sizes.

In an effort to address watershed scale forecast skill, watershed intensity error as defined earlier was investigated for three key watersheds in California: the Russian, Yuba/Feather and Santa Ana River watersheds (Fig. 9). In the Russian and Yuba/Feather watersheds, both models show clear improvement with the control forecasts, both in terms of counts of instances where the control forecast has less error (Figs. 9a,b) and in terms of the average improvement of those instances (Figs. 9c,d). Over the Santa Ana River watershed the control forecast showed more instances with less error in the ECMWF model, but a larger mean difference in the denial forecasts, and results were relatively neutral for the NCEP model. Despite the lack of a clear signal over the Santa Ana River watershed, this analysis suggests that the dropsondes did in general have a positive influence on the precipitation forecasts at the watershed scale over California.

4. Discussion

Based on data denial experiments, dropsondes from AR Recon often added value to the forecasts of both IVT and precipitation for both the NCEP and ECMWF global models during 2019 and 2020, as summarized in Fig. 10. For comparisons in the figure between the control and denial that are based on a count of instances or a mean value (i.e., Figs. 5 and 9), we have subjectively chosen to label one better than the other if the difference in the count or mean is greater than 10% of the value of the control. While there are several cases where the dropsondes did not add value, the overall influence of the dropsondes is a clear improvement. Based on Lorenc and Marriott (2014), the strength of the signal of improvement would likely be clearer with a larger number of IOPs in the analysis.

The metrics used in this study emphasize different attributes of skill to present a more complete picture of the influences of dropsonde observations on the forecasts. MAE is a measure of the magnitude of the errors, while the correlation is measure of the error in the spatial pattern of IVT or precipitation. Both MAE and correlation are point-based metrics which do not consider neighboring grid points, while FSS and watershed intensity look at precipitation over larger areas. For MAE and correlation, we considered significance for individual forecasts using SPCT, which highlights single forecasts where the dropsondes had the greatest influence. Basing significance on averages over many forecasts considers the improvement in a broader sense.

Improvement across different metrics for the control forecasts points to the robustness of the improvement. For example, the enhancements in precipitation skill for both point-based metrics and metrics covering larger areas suggests that the improvement in precipitation skill occurs across spatial scales. Conversely, differences between results can also provide useful information. For example, there were largely consistent increases in IVT correlation, but there were fewer reductions in IVT MAE. This suggests that the dropsondes aid more in the positioning of an AR than with the magnitude. The opposite was true for precipitation, where MAE was generally improved while correlation often was not. This suggests an improvement in precipitation bias, without an improvement in location.

There are several instances when either the IVT or the precipitation demonstrated improvements with the dropsondes, but not both fields. In some cases, the precipitation is improved 24 h after the IVT improvement. Since the IVT is a "snapshot" every 24 h, while the precipitation is accumulated over 24 h, it is reasonable that in some instances precipitation improvement would correspond to the IVT improvement at the beginning of the accumulation and not the end. IVT integrated over 24 h would likely have a stronger correspondence to 24-h accumulated precipitation than an IVT "snapshot." Another difference is that the IVT verification is computed over both land and ocean, while the precipitation verification is limited to land only, which is both a more limited domain and farther from the locations of the dropsondes. More generally, precipitation and IVT could have different responses since they are very different parameters with different characteristics and predictability. While IVT is a necessary factor for predicting precipitation (Lavers et al. 2016), there are multiple other factors, such as local terrain, model resolution, vertical motion, AR orientation, and microphysics, that influence the forecast precipitation accumulation.

There were also differences in the influences of dropsondes between the two models. These differences could be due to many factors, including the data assimilation details (e.g., the analysis scheme, the background error covariance matrix, initial imbalance, handling of observed time information, or the amount of satellite data assimilated), as well as the model resolution, and model physical schemes. Earlier literature on adaptive observations (e.g., Bergot 1999; Weissmann et al. 2011) pointed out that the data assimilation method often has a substantial impact on the data impact of adaptive observations with an NWP system. Of particular note is use of a 12-h time window for the ECMWF model data assimilation compared to the use of a 6-h time window for the NCEP model data assimilation, which could result in smoother fields

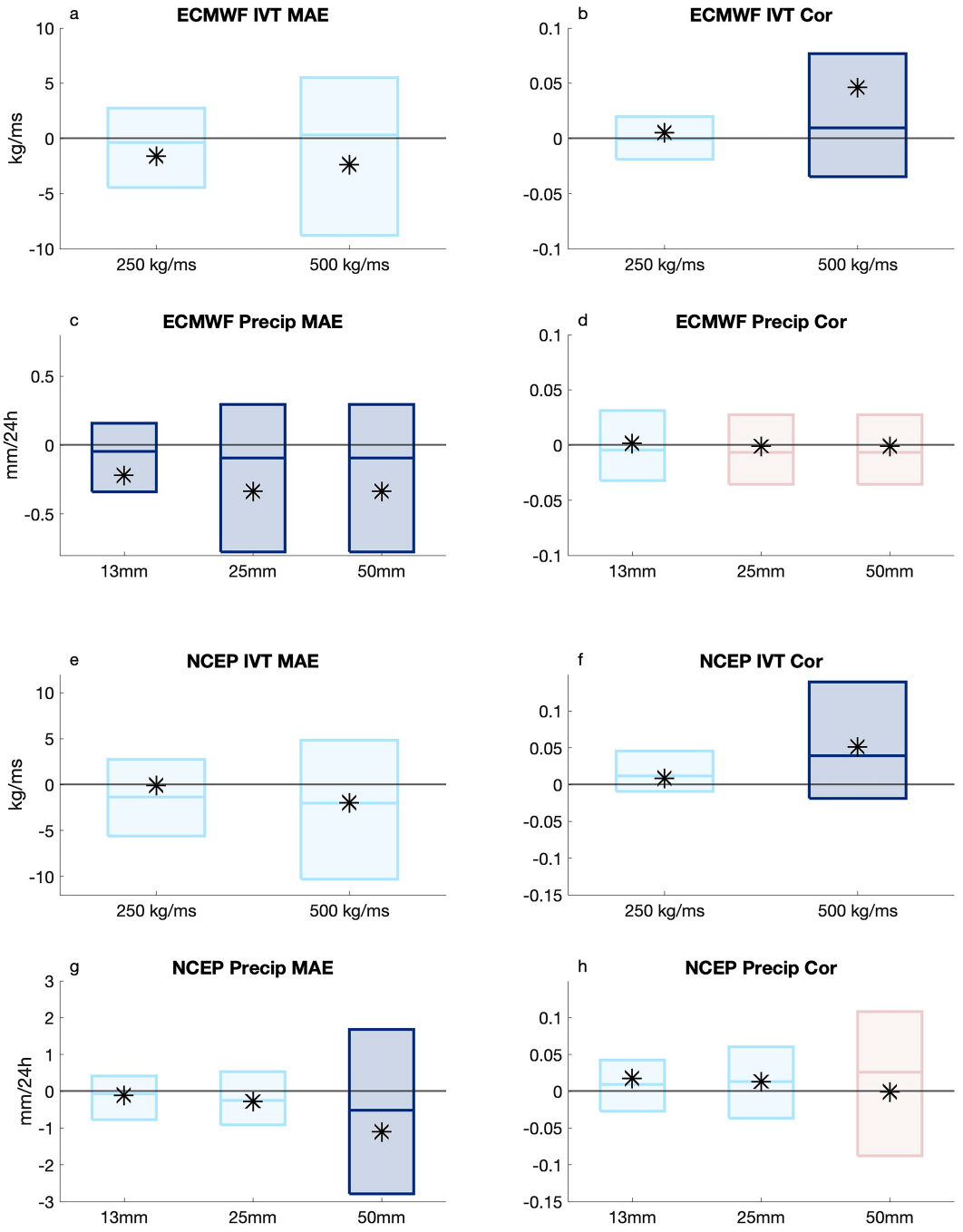

FIG. 7. As in Fig. 6, but averaged over lead times rather than thresholds.

assimilated into the ECMWF model and more small-scale features for the NCEP model in short-range forecasts. However, both models were running in cycled systems, therefore, it is almost impossible to quantify what factors are dominating the model differences, and that is outside the scope of this study. Generally, it is considered that the quality of a model's first guess and the capability of the observations in filling gaps left from other observations are critical in assessing additional impacts from target observations (Majumdar 2016; Zheng et al. 2021b; Sun et al. 2022).

The NCEP model showed more improvement in precipitation forecasts with the dropsondes at shorter lead times than at longer lead times. Meanwhile, the positive impact on IVT averaged at all lead times is more pronounced than on the precipitation. These results are qualitatively consistent with the findings in Zheng et al. (2021b), which utilized a regional

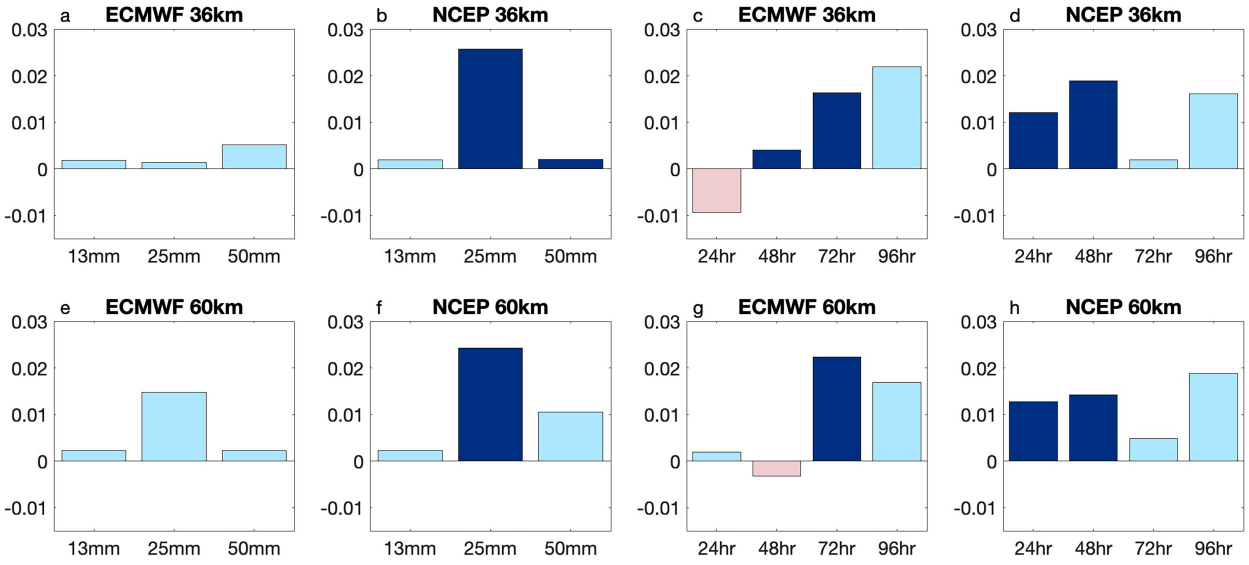

FIG. 8. Difference in fractions skill score (FSS) between control and denial for both the ECMWF and NCEP models, (a),(b),(e),(f) averaged over all lead times and (c),(d),(g),(h) averaged over all thresholds using a (top) 36-km neighborhood and (bottom) a 60-km neighborhood. Blue colors indicate that the control has a higher score, and red colors indicate that the denial has a higher score. Dark colors indicate significance based on a 90% confidence interval computed with bootstrapping.

modeling system forced by the NCEP GFS analysis and forecast products. In addition, the NCEP model has most precipitation skill improvements at the higher thresholds, in agreement with Lord et al. (2023a), while the ECMWF precipitation results were less sensitive to thresholds. For the ECMWF model, the control forecasts have less skill than the denial at a 24-h lead time but show consistent and generally significant improvement at longer lead times. One possible factor contributing to the

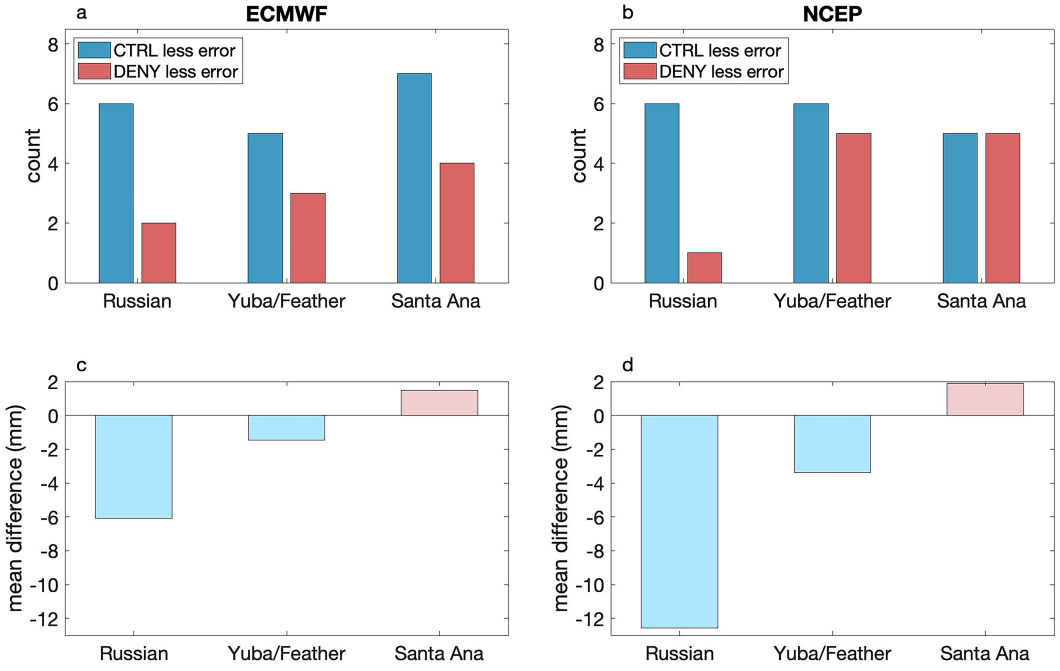

FIG. 9. Counts of instances (valid days and lead times) where each of the control and denial forecasts had (a),(b) a smaller watershed intensity error magnitude for three California watersheds and (c),(d) the mean difference (control − denial) of the magnitude of the error for those instances. The counts are limited to cases where the difference between control and denial is greater than 1 mm $(24\ \text{h})^{-1}$.

	ECMWF IVT	ECMWF precip	NCEP IVT	NCEP precip
	count / mean	count / mean	count / mean	count / mean
MAE SPCT	● / ●	● / ●	● / ✕	✕ / ●
Correlation SPCT	● / ●	✕ / ●	● / ✕	– / –
Watershed: Russian		● / ●		● / ●
Watershed: Yuba/Feather		● / ●		● / ●
Watershed: Santa Ana		● / ✕		– / ✕
by lead time	24h 48h 72h	24h 48h 72h 96h	24h 48h 72h	24h 48h 72h 96h
MAE	– – –	✕ ● ● ●	– – –	● – ● –
Correlation	● – ●	✕ – – –	● ● ●	– ● ● –
FSS (36km)		– ● ● –		● ● – –
FSS (60km)		– – ● –		● ● – –
by threshold	250 500 (kg/ms)	13 25 50 (mm)	250 500 (kg/ms)	13 25 50 (mm)
MAE	– –	● ● ●	– –	– – ●
Correlation	– ●	– – –	– ●	– – –
FSS (36km)		– – –		– ● ●
FSS(60km)		– – –		– ● –

FIG. 10. Summary of control and denial experiment results. Blue circles indicate that the control is better, red X marks indicate that the denial is better, and dashes indicate neither outperforms the other. For metrics by count and mean, a dash is used when the difference is less than 10% of the control. For metrics by lead time or threshold, symbols are shown for each lead time or threshold in increasing order [24, 48 h etc., or 13, 25 mm $(24\ \text{h})^{-1}$, etc.], and dashes are used when the differences are not significant at 90%.

reduction in skill for the 24-h forecast is that the relative high skill of the denial forecast at the 24-h lead time (compared to longer lead times) results in less improvement with the addition of the dropsondes.

5. Conclusions

We set out to investigate whether the targeted sampling done as part of Atmospheric River Reconnaissance (AR Recon) resulted in quantitative improvements of integrated vapor transport (IVT) and precipitation forecasts in two global models, according to AR Recon's stated goals. The dropsonde impacts were extensively investigated during the AR Recon 2019 and 2020 periods with the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) and the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) models.

Our work illustrates that, more often than not, forecasts were improved across 22 different IOPs in 2019 and 2020 when dropsonde data were assimilated. Both the ECMWF IFS and the NCEP GFS models show many improvements in forecast skill with the added information from the dropsondes, which aligns well with the work discussed in the Introduction. In particular, significant improvements in the control forecast IVT generally occur in both models, especially at a higher threshold. Significant improvements in the control forecast precipitation also generally occur in both models, but the two models are not consistent in the lead times and metrics that demonstrate the improvements.

AR Recon flight planning strategies are continually adapting and improving based on the latest science. One key example of this is the development of AR Recon sequences, consecutive days with flights, based on the results of Zheng et al. (2021b) and Stone et al. (2020). The results from this study will further aid in the development and implementation of impact-informed sampling strategies, helping flight planners target specific areas that should have the greatest impact on the forecast. Other future work may investigate the model sensitivity illustrated by the differences between the models. The differences between the IFS and GFS responses to dropsonde assimilation may lead to physical insights into model parameterization, if there are identifiable patterns, which may be explored more readily as we collect more samples of AR storms and model runs with and without the dropsonde data. Future work will also focus on

assessing data impacts under the framework of probabilistic forecasting.

Acknowledgments. This research was funded by the U.S. Army Corps of Engineers (USACE) as part of Forecast Informed Reservoir Operations under Grant W912HZ-15-2-0019 and by the California Department of Water Resources Atmospheric River Program (4600013361). David Lavers was supported by the Copernicus Climate Change Service, which is implemented by ECMWF on behalf of the European Union.

Data availability statement. The data that support the findings of this study will be openly available from the University of California, San Diego, Library Digital Collections (https://doi.org/10.6075/J0RJ4JPW).

REFERENCES

Arnold, C. P., Jr., and C. H. Dey, 1986: Observing-systems simulation experiments: Past, present, and future. *Bull. Amer. Meteor. Soc.*, **67**, 687–695, https://doi.org/10.1175/1520-0477(1986)067<0687:OSSEPP>2.0.CO;2.

Beck, H. E., and Coauthors, 2019: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrol. Earth Syst. Sci.*, **23**, 207–224, https://doi.org/10.5194/hess-23-207-2019.

Bergot, T., 1999: Adaptive observations during FASTEX: A systematic survey of upstream flights. *Quart. J. Roy. Meteor. Soc.*, **125**, 3271–3298, https://doi.org/10.1002/qj.49712556108.

Bullock, R. G., B. G. Brown, and T. L. Fowler, 2016: Method for object-based diagnostic evaluation. NCAR Tech. Note NCAR/TN-532+STR, 84 pp., https://doi.org/10.5065/D61V5CBS.

Cardinali, C., 2009: Monitoring the observation impact on the short-range forecast. *Quart. J. Roy. Meteor. Soc.*, **135**, 239–250, https://doi.org/10.1002/qj.366.

Centurioni, L., A. Horányi, C. Cardinali, E. Charpentier, and R. Lumpkin, 2017: A global ocean observing system for measuring sea level atmospheric pressure: Effects and impacts on numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **98**, 231–238, https://doi.org/10.1175/BAMS-D-15-00080.1.

Chen, S., C. A. Reynolds, J. M. Schmidt, P. P. Papin, M. A. Janiga, R. Bankert, and A. Huang, 2022: The effect of a Kona low on the eastern Pacific Valentine's Day (2019) atmospheric river. *Mon. Wea. Rev.*, **150**, 863–882, https://doi.org/10.1175/MWR-D-21-0182.1.

Cobb, A., L. Delle Monache, F. Cannon, and F. M. Ralph, 2021: Representation of dropsonde-observed atmospheric river conditions in reanalyses. *Geophys. Res. Lett.*, **48**, e2021GL093357, https://doi.org/10.1029/2021GL093357.

——, and Coauthors, 2023: Atmospheric river reconnaissance 2021: A review. *Wea. Forecasting*, https://doi.org/10.1175/WAF-D-21-0164.1, in press.

Cordeira, J. M., F. M. Ralph, A. Martin, N. Gaggini, J. R. Spackman, P. J. Neiman, J. J. Rutz, and R. Pierce, 2017: Forecasting atmospheric rivers during CalWater 2015. *Bull. Amer. Meteor. Soc.*, **98**, 449–459, https://doi.org/10.1175/BAMS-D-15-00245.1.

DeHaan, L. L., A. C. Martin, R. R. Weihs, L. Delle Monache, and F. M. Ralph, 2021: Object-based verification of atmospheric river predictions in the northeast Pacific. *Wea. Forecasting*, **36**, 1575–1587, https://doi.org/10.1175/WAF-D-20-0236.1.

Demirdjian, R., J. D. Doyle, C. A. Reynolds, J. R. Norris, A. C. Michaelis, and F. M. Ralph, 2020: A case study of the physical processes associated with the atmospheric river initial-condition sensitivity from an adjoint model. *J. Atmos. Sci.*, **77**, 691–709, https://doi.org/10.1175/JAS-D-19-0155.1.

Doyle, J. D., C. Amerault, C. A. Reynolds, and P. A. Reinecke, 2014: Initial condition sensitivity and predictability of a severe extratropical cyclone using a moist adjoint. *Mon. Wea. Rev.*, **142**, 320–342, https://doi.org/10.1175/MWR-D-13-00201.1.

Geer, A. J., and Coauthors, 2021: Bulk hydrometeor optical properties for microwave and sub-millimetre radiative transfer in RTTOV-SCATT v13.0. *Geosci. Model Dev.*, **14**, 7497–7526, https://doi.org/10.5194/gmd-14-7497-2021.

Gilleland, E., 2013: Testing competing precipitation forecasts accurately and efficiently: The spatial prediction comparison test. *Mon. Wea. Rev.*, **141**, 340–355, https://doi.org/10.1175/MWR-D-12-00155.1.

——, D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Haase, J. S., M. J. Murphy, B. Cao, F. M. Ralph, M. Zheng, and L. Delle Monache, 2021: Multi-GNSS airborne radio occultation observations as a complement to dropsondes in atmospheric river reconnaissance. *J. Geophys. Res. Atmos.*, **126**, e2021JD034865, https://doi.org/10.1029/2021JD034865.

Harris, L. M., and S.-J. Lin, 2013: A two-way nested global-regional dynamical core on the cubed-sphere grid. *Mon. Wea. Rev.*, **141**, 283–306, https://doi.org/10.1175/MWR-D-11-00201.1.

——, L. Zhou, X. Chen, and J.-H. Chen, 2020: The GFDL finite-volume cubed-sphere dynamical core: Release 201912. NOAA Tech. Memo. OAR GFDL2020-001, 10 pp., https://doi.org/10.25923/7h88-c534.

Hatchett, B. J., and Coauthors, 2020: Observations of an extreme atmospheric river storm with a diverse sensor network. *Earth Space Sci.*, **7**, e2020EA001129, https://doi.org/10.1029/2020EA001129.

Hering, A. S., and M. G. Genton, 2011: Comparing spatial predictions. *Technometrics*, **53**, 414–425, https://doi.org/10.1198/TECH.2011.10136.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Ingleby, B., L. Isaksen, T. Kral, T. Haiden, and M. Dahoui, 2018: Improved use of atmospheric in situ data. *ECMWF Newsletter*, No. 155, ECMWF, Reading, United Kingdom, 9 pp., https://www.ecmwf.int/en/newsletter/155/meteorology/improved-use-atmospheric-situ-data.

——, F. Prates, L. Isaksen, and M. Bonavita, 2019: Recent BUFR dropsonde data improved forecasts. *ECMWF Newsletter*, No. 162, ECMWF, Reading, United Kingdom, 9 pp., https://www.ecmwf.int/en/newsletter/162/news/recent-bufr-dropsonde-data-improved-forecasts.

Jasperse, J., and Coauthors, 2020: Lake Mendocino forecast informed reservoir operations: Final viability assessment. UC San Diego, 142 pp., https://escholarship.org/uc/item/3b63q04n.

Kleist, D. T., and K. Ide, 2015: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Wea. Rev.*, **143**, 452–470, https://doi.org/10.1175/MWR-D-13-00350.1.

Lavers, D. A., D. E. Waliser, F. M. Ralph, and M. D. Dettinger, 2016: Predictability of horizontal water vapor transport relative to precipitation: Enhancing situational awareness for

forecasting western U.S. extreme precipitation and flooding. *Geophys. Res. Lett.*, **43**, 2275–2282, https://doi.org/10.1002/2016GL067765.

——, M. J. Rodwell, D. S. Richardson, F. M. Ralph, J. D. Doyle, C. A. Reynolds, V. Tallapragada, and F. Pappenberger, 2018: The gauging and modeling of rivers in the sky. *Geophys. Res. Lett.*, **45**, 7828–7834, https://doi.org/10.1029/2018GL079019.

Lin, S.-J., 2004: A "vertically Lagrangian" finite-volume dynamical core for global models. *Mon. Wea. Rev.*, **132**, 2293–2307, https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2.

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2, https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.

Lord, S. J., X. Wu, V. Tallapragada, and F. M. Ralph, 2023a: The impact of dropsonde data on the performance of the NCEP Global Forecast System during the 2020 atmospheric rivers observing campaign. Part I: Precipitation. *Wea. Forecasting*, **38**, 17–45, https://doi.org/10.1175/WAF-D-22-0036.1.

——, ——, ——, and ——, 2023b: The impact of dropsonde data on the performance of the NCEP Global Forecast System during the 2020 atmospheric rivers observing campaign. Part II: Dynamic variables and humidity. *Wea. Forecasting*, **38**, 721–752, https://doi.org/10.1175/WAF-D-22-0072.1.

Lorenc, A. C., and R. T. Marriott, 2014: Forecast sensitivity to observations in the Met Office global numerical weather prediction system. *Quart. J. Roy. Meteor. Soc.*, **140**, 209–224, https://doi.org/10.1002/qj.2122.

Majumdar, S. J., 2016: A review of targeted observations. *Bull. Amer. Meteor. Soc.*, **97**, 2287–2303, https://doi.org/10.1175/BAMS-D-14-00259.1.

Neiman, P. J., B. J. Moore, A. B. White, G. A. Wick, J. Aikins, D. L. Jackson, J. R. Spackman, and F. M. Ralph, 2016: An airborne and ground-based study of a long-lived and intense atmospheric river with mesoscale frontal waves impacting California during CalWater-2014. *Mon. Wea. Rev.*, **144**, 1115–1144, https://doi.org/10.1175/MWR-D-15-0319.1.

Pauley, P. M., and B. Ingleby, 2022: Assimilation of in-situ observations. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, S. K. Park and L. Xu, Eds., Springer, 293–371, https://link.springer.com/book/10.1007/978-3-030-77722-7.

Putman, W. M., and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *J. Comput. Phys.*, **227**, 55–78, https://doi.org/10.1016/j.jcp.2007.07.022.

Rabier, F., E. Klinker, P. Courtier, and A. Hollingsworth, 1996: Sensitivity of forecast errors to initial conditions. *Quart. J. Roy. Meteor. Soc.*, **122**, 121–150, https://doi.org/10.1002/qj.49712252906.

——, H. Jarvinen, E. Klinker, J.-F. Mahfouf, and A. Simmons, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quart. J. Roy. Meteor. Soc.*, **126**, 1143–1170, https://doi.org/10.1002/qj.49712656415.

Ralph, F. M., and Coauthors, 2014: A vision for future observations for western U.S. extreme precipitation and flooding. *J. Contemp. Water Res. Educ.*, **153**, 16–32, https://doi.org/10.1111/j.1936-704X.2014.03176.x.

——, and Coauthors, 2017: Dropsonde observations of total integrated water vapor transport within North Pacific atmospheric rivers. *J. Hydrometeor.*, **18**, 2577–2596, https://doi.org/10.1175/JHM-D-17-0036.1.

——, J. J. Rutz, J. M. Cordeira, M. Dettinger, M. Anderson, D. Reynolds, L. J. Schick, and C. Smallcomb, 2019: A scale to characterize the strength and impacts of atmospheric rivers. *Bull. Amer. Meteor. Soc.*, **100**, 269–289, https://doi.org/10.1175/BAMS-D-18-0023.1.

——, and Coauthors, 2020: West Coast forecast challenges and development of atmospheric river reconnaissance. *Bull. Amer. Meteor. Soc.*, **101**, E1357–E1377, https://doi.org/10.1175/BAMS-D-19-0183.1.

Reynolds, C. A., J. D. Doyle, F. M. Ralph, and R. Demirdjian, 2019: Adjoint sensitivity of North Pacific atmospheric river forecasts. *Mon. Wea. Rev.*, **147**, 1871–1897, https://doi.org/10.1175/MWR-D-18-0347.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Skamarock, W. C., and Coauthors, 2019: A description of the Advanced Research WRF Model version 4. NCAR Tech. Note NCAR/TN-556+STR, 145 pp., https://doi.org/10.5065/1dfh-6p97.

Stone, R. E., C. A. Reynolds, J. D. Doyle, R. H. Langland, N. L. Baker, D. A. Lavers, and F. M. Ralph, 2020: Atmospheric river reconnaissance observation impact in the Navy global forecast system. *Mon. Wea. Rev.*, **148**, 763–782, https://doi.org/10.1175/MWR-D-19-0101.1.

Sun, W., Z. Liu, C. A. Davis, F. M. Ralph, L. Delle Monache, and M. Zheng, 2022: Impacts of dropsonde and satellite observations on the forecasts of two atmospheric-river-related heavy rainfall events. *Atmos. Res.*, **278**, 106327, https://doi.org/10.1016/j.atmosres.2022.106327.

Szunyogh, I., Z. Toth, R. E. Morss, S. J. Majumdar, B. J. Etherton, and C. H. Bishop, 2000: The effect of targeted dropsonde observations during the 1999 winter storm reconnaissance program. *Mon. Wea. Rev.*, **128**, 3520–3537, https://doi.org/10.1175/1520-0493(2000)128<3520:TEOTDO>2.0.CO;2.

Torn, R. D., and G. J. Hakim, 2008: Ensemble-based sensitivity analysis. *Mon. Wea. Rev.*, **136**, 663–677, https://doi.org/10.1175/2007MWR2132.1.

——, and ——, 2009: Initial condition sensitivity of western Pacific extratropical transitions determined using ensemble-based sensitivity analysis. *Mon. Wea. Rev.*, **137**, 3388–3406, https://doi.org/10.1175/2009MWR2879.1.

Wang, X., and T. Lei, 2014: GSI-based four-dimensional ensemble–variational (4DEnsVar) data assimilation: Formulation and single-resolution experiments with real data for NCEP global forecast system. *Mon. Wea. Rev.*, **142**, 3303–3325, https://doi.org/10.1175/MWR-D-13-00303.1.

Weissmann, M., and Coauthors, 2011: The influence of assimilating dropsonde data on typhoon track and midlatitude forecasts. *Mon. Wea. Rev.*, **139**, 908–920, https://doi.org/10.1175/2010MWR3377.1.

Yang, F., and V. Tallapragada, 2018: Evaluation of retrospective and real-time NGGPS FV3GFS experiments for Q3FY18 beta implementation. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 5B.3, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345231.html.

Zheng, M., E. K. M. Chang, and B. A. Colle, 2013: Ensemble sensitivity tools for assessing extratropical cyclone intensity and

track predictability. *Wea. Forecasting*, **28**, 1133–1156, https://doi.org/10.1175/WAF-D-12-00132.1.

——, and Coauthors, 2021a: Data gaps within atmospheric rivers over the northeastern Pacific. *Bull. Amer. Meteor. Soc.*, **102**, E492–E524, https://doi.org/10.1175/BAMS-D-19-0287.1.

——, and Coauthors, 2021b: Improved forecast skill through the assimilation of dropsonde observations from the atmospheric river reconnaissance program. *J. Geophys. Res. Atmos.*, **126**, e2021JD034967, https://doi.org/10.1029/2021JD034967.

Zhou, L., S.-J. Lin, J.-H. Chen, L. M. Harris, X. Chen, and S. L. Rees, 2019: Toward convective-scale prediction within the next generation global prediction system. *Bull. Amer. Meteor. Soc.*, **100**, 1225–1243, https://doi.org/10.1175/BAMS-D-17-0246.1.